You talking to me? A predictive model for zero auxiliary constructions

Andrew Caines
Computation, Cognition & Language Group
RCEAL, University of Cambridge, UK
apc38@cam.ac.uk

Paula Buttery
Computation, Cognition & Language Group
RCEAL, University of Cambridge, UK
pjb48@cam.ac.uk

Abstract

As a consequence of the established practice to prefer training data obtained from written sources, NLP tools encounter problems in handling data from the spoken domain. However, accurate models of spoken data are increasingly in demand for naturalistic speech generation and machine translations in speech-like contexts (such as chat windows and SMS). There is a widely held assumption in the linguistic field that spoken language is an impoverished form of written language. However, we show that spoken data is not unpredictably irregular and that language models can benefit from detailed consideration of spoken language features. This paper considers one specific construction which is largely restricted to the spoken domain - the ZERO AUXILIARY - and makes a predictive model of that construction for native speakers of British English. The model can predict zero auxiliary occurrence in the BNC with 96.9% accuracy. We will demonstrate how this model can be integrated into existing parsing tools, increasing the number of successful parses for this zero auxiliary construction by around 30%, and thus improving the performance of NLP applications which rely on parsing.

1 Introduction

Up to this point, statistical Natural Language Processing (NLP) tools have generally been trained on corpora that are representative of written rather than spoken language. A major factor behind this decision to use written data is that it is far easier to collect than spoken data. Newswire, for instance, may be harvested readily and in abundance. Once collected, written language requires relatively little processing before it can be used for training a statistical model.

Processing of spoken data, on the other hand, involves at the very least transcription - which usually requires a human transcriber. Since transcription is a slow and laborious task, the collection of spoken data is highly resource intensive. But this relative difficulty in collection is not the only reason that spoken language data has been sidelined. Had spoken data been considered to be crucial to the production of NLP applications greater efforts might have been made to obtain it. However, on account of some of its characteristic features such as hesitations, interruptions and ellipsis, spoken language is often dismissed as nothing more than a noisy approximation to 'real' or 'intended' language.

In some forums, written language is held up as an idealised form of language toward which speakers aspire and onto which spoken language should be retrofitted. This is an artefact of the theoretical notion of a 'competence'-'performance' dichotomy (Chomsky 1965) with the latter deemed irrelevant and ignored in mainstream linguistic research.

The consequence of the established practice to sideline spoken data is that NLP tools are inherently error prone when handling data from the spoken domain. With increasing calls for speech to be considered the primary form of language and to be treated as such (Sampson 2001: 7 [1]; Cermák 2009: 115 [2]; Haugh 2009: 74 [3]) and a growing trend for NLP techniques to be integrated into cognitive and neurolinguistic research as well as forensic appli-

[1] Speech is "unquestionably the more natural, basic mode of language behaviour".

[2] "From a linguistic point of view, spoken corpora should be primary for research but that has not been the case so far".

[3] Haugh observes that "spoken language and interaction lie at the core of human experience" but bemoans the "relative neglect of spoken language in corpora to date".

Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, ACL 2010, pages 43–51,
Uppsala, Sweden, 16 July 2010.

cations, there are now compelling reasons to examine spoken data more closely. Accurate models of spoken data are increasingly in demand for naturalistic speech generation and machine translations in speech-like contexts (such as human-machine dialogue, chat windows and SMS).

The main research aim of our work is to show that spoken data should not be considered error prone and therefore unpredictably irregular. We show that language models can be improved in increments as we deepen our understanding of spoken language features. We investigate ZERO AUXILIARY progressive aspect constructions - those which do not feature the supposedly obligatory auxiliary verb, as in (1a) below (cf. 1b):

(1a) What you doing? Who you looking for? You been working?

(1b) What are you doing? Who are you looking for? Have you been working?

The zero auxiliary is a non-standard feature which for the most part is known to be restricted to speech. A corpus study of spoken British English indicates that in progressive aspect interrogatives with second person subjects (as in (1) above) the auxiliary occurs in zero form in 27% of constructions found. The equivalent figure from the written section of the corpus is just 5.4%. Consequently, existing NLP techniques - since they are based on written training data - are unlikely to deal appropriately with zero auxiliary constructions. We report below on the corpus study in full and use the results of logistic regression to design a predictive model of zero auxiliary occurrence in spoken English. The model is based on contextual grammatical features and can predict zero auxiliary occurrence in the British National Corpus (BNC; 2007) with 96.9% accuracy. Finally, we discuss how this model can be used to improve the performance of NLP techniques in the spoken domain, demonstrating its implementation in the RASP system (Robust Accurate Statistical Parsing; (Briscoe, Carroll and Watson, 2006)).

This paper underlines why awareness of non-standard linguistic features matters. Targeted data extraction from large corpus resources allows the construction of more informed language models which have been trained on naturalistic spoken usage rather than standard and restricted rules of written language. Such work has only been made possible with the advent of large spoken language corpora such as the BNC. Even so, the resource-heavy nature of spoken data collection means that speech transcriptions constitute only one tenth of this 100 million word corpus [4]. Nevertheless, it is an invaluable resource made up of a range of speech genres including spontaneous face-to-face conversation, a fact which makes it unique among corpora. Since conversational dialogue is the predominant language medium, the BNC offers the best chance of modelling speech as it occurs naturally.

This work has important implications for both computational and theoretical linguistics. On the one hand, we can improve various NLP techniques with more informed language models, and on the other hand we are reminded that the space of grammatical possibility is not restricted and that continued empirical investigation is key in order to arrive at the fullest possible description of language use.

2 Spoken and written language

In the modern mainstream fields of linguistic research, based on Chomsky's 'ideal speaker-listener' (1965), spoken language has been all too easily dismissed from consideration on the grounds that it is more error-prone and less important than written language. In this idealisation, the speaker-listener is "unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic)" (Chomsky, 1965).

The 'errors' Chomsky refers to are features of speech production such as pauses, filled silence, hesitation, repetition and elision, or of dialogue such as backgrounding, overlap and truncation between speakers. Thus 'error' is essentially here defined as that which is not normally found in well-formed written data, that which is 'noisy' and 'unpredictable'. It is on these grounds - the grammatical rigidity of the written medium relative to speech - that the divide between spoken and written language modelling has grown up.

The opposing, usage-based view is that spoken language is systematic and that it should be modelled as it is rather than as a crude approximation of the written form. On this view, the speech production and dialogue features listed above are not

[4] Cermák estimates our experience with each language medium is in fact this ratio in reverse - 90:10 spoken to written (2009: 115).

considered mistakes but "regular products of the system of spoken English" (Halliday, 1994). 'Error' is thus seen as a misnomer for these features because they are in fact all regular in some way. For example, the fact that people tend to put filler pauses in specific places.

We propose a middle way: that which builds on the NLP tools available, even though they are trained on written data, and on top of these models the features of spoken language as 'noise' in the communicative channel. This is a pragmatic approach which recognises the considerable amount of work underpinning existing NLP tools and sees no value in discarding these and starting again from scratch. We demonstrate that spoken language is model-able and predictable, even with a feature which would not be seen as 'correct' in written form. For practical purposes we need to recognise the regularities in the apparently 'incorrect' features of speech and build these into the functioning language models we already have through statistical analysis of corpus distributions and appropriate adjustment to parser tools.

3 The zero auxiliary construction

According to standard grammatical rules, the auxiliary verb is an obligatory feature of progressive aspect constructions. However, this rule is based on norms of written language and is in fact not always adhered to in the production of speech. As a result, some progressive constructions do not feature an auxiliary verb. These are termed 'zero auxiliary' constructions and have been previously examined in studies of dialect (Labov, 1969; Andersen, 1995) and first language acquisition (Brown, 1973; Rizzi, 1993/1994; Wexler, 1994; Lieven et al, 2003; Wilson, 2003; Theakston et al, 2005).

There are copious anecdotal examples of the zero auxiliary:

(2) You talking to me? Travis Bickle in *Taxi Driver* (1976).

(3) Where he going? Avon Barksdale in *The Wire*, Season 1: 'Game Day' (2002).

(4) What you doing? Holly Golightly in *Breakfast at Tiffany's* (1961).

Natural language data taken from the spoken section of the British National Corpus (sBNC) shows that the zero auxiliary features in 1330 (27%) of the 4923 second person progressive interrogative constructions; as in (1), (2), (4) above. In first person singular declaratives (cf. (5a) and (5b)), in contrast, the proportion of zero auxiliary occurrence is just 0.9% (158 of 17,838 constructions). This already indicates the way that the zero auxiliary occurs in predictable contexts and how grammatical properties will feature in the predictive model.

(5a) What I saying? I annoying you? Why I doing this?

(5b) What am I saying? Am I annoying you? Why am I doing this?

Subject person, subject number, subject type (pronoun or other noun) and clause type (declarative or interrogative) are four of the eight syntactic properties incorporated in the predictive model. The four other properties are clause tense (6), perfect or non-perfect aspect (7), clause polarity (8) and presence or absence of subject (9).

(6) You are debugging. You were debugging.

(7) We have been looking for a present. We are looking for a present.

(8) She is watching the grand prix. He is not watching the grand prix.

(9) I am going to town in a minute. Going to town in a minute.

We employ logistic regression to investigate the precise nature of the relationships between zero auxiliary use and these various linguistic variables. This allows us to build a predictive model of zero auxiliary occurrence in spoken language which will be useful for several reasons relating to parsing of natural spoken language. Firstly, for automatic parsing of spoken data being able to predict when a zero auxiliary is likely to occur enables the parser to relax its normal rules which are based on written standards. Secondly, as technology improves and interaction with computers becomes more humanistic the need to replicate human-like communication increases in importance: by knowing in which contexts the auxiliary verb might be absent, researchers can build a language model which is more realistic and so the user experience is improved and made more naturalistic. Thirdly, a missing auxiliary might be problematic for machine translation since it could

result in the loss of tense and aspect information, but with the ability to predict where a zero auxiliary might occur, the auxiliary can be restored so that translation can be performed with appropriate tense and aspect.

For all these reasons, the zero auxiliary in spoken English is an appropriate case study for finding the common ground between NLP and Linguistics. Awareness of this particular linguistic phenomenon through corpus study allows the construction of more informed language models which in turn enhance relevant NLP techniques. The cross-pollination of research from NLP and linguistics benefits both fields and ties in with the emergence of linguistic theories that "conceive of structure as gradient, malleable and probabilistic" and incorporate "knowledge of the frequency and probability of use of these categories in speakers' experience" (Tily et al, 2009). These are collectively known as 'usage-based' approaches to language theory and are exerting a growing influence on the field (e.g. Barlow and Kemmer 2000; Bybee and Hopper 2001; Bod, Hay and Jannedy 2003).

4 Corpus study

Training data was obtained through manual annotation of progressive constructions in the British National Corpus (2007). A preliminary study of interrogatives with second person subjects confirmed that the zero auxiliary is more a feature of the spoken rather than the written domain (Table 1). Therefore a focus on the spoken section of the corpus (sBNC) was justified and so we undertook a comprehensive study of all progressive constructions in sBNC. The genres contained in sBNC include a range of settings and levels of formality - from academic lectures to radio programmes to spontaneous, face-to-face conversation.

We extracted 93,253 sentences featuring a progressive construction from sBNC and each was manually annotated for auxiliary realisation and the eight syntactic properties described in Table 2. In Table 3 the progressive constructions are classified by auxiliary realisation. With approximately 4.2% occurrence in progressive constructions, zero auxiliaries are a low frequency feature of spoken language but ones which are significant for the fact that existing NLP tools cannot successfully parse them, thus one in twenty-five progressive constructions will not be fully parsed. We use the annotated corpus of these progressive constructions to design the predictive model described below.

Corpus	**Auxiliary**		
	Full	**Contracted**	**Zero**
wBNC	3220	27	187
sBNC	3498	95	1330

Table 1: Auxiliary realisation in second person progressive interrogatives in the BNC.

Properties	Value encodings
Aux realisation	
Zero auxiliary	full(0), contracted(1), zero(2)
Variables	
Subject person	1st(1), 2nd(2), 3rd(3)
Subject number	singular(0), plural(1)
Subject type	other noun(0), pronoun(1)
Subj supplied	zero subj(0), subj supplied(1)
Clause type	declarative(0), interrogative(1)
Clause tense	present(0), past(1)
Perfect aspect	non-perfect(0), perfect(1)
Polarity	positive(0), negative(1)

Table 2: Syntactic features and their encodings in the annotated sBNC Progressive Corpus

5 Model

To predict the zero auxiliary in spoken language we use logistic regression. To train this model we took a 90% sample from our corpus of 93,253 progressive constructions extracted from the spoken section of the BNC, as described above and in Caines 2010. The dataset was split into two categories: those sentences which exhibited the zero auxiliary and those which did not[5]. A logistic regression was then performed to ascertain the probability of category membership using the eight previously described syntactic properties. Note that subject person is arguably not a scalar vari-

[5]Contracted auxiliaries thus belong in the 'not zero auxiliary' category.

Corpus	**Full**	**Contracted**	**Zero**
sBNC	38,015	51,295	3943

Table 3: Auxiliary realisation in progressive constructions in sBNC.

able and therefore is re-analysed as three boolean variables with separate binary values for use of the first, second and third person. However, the three subject person variables are dependent (*ie*. If the subject is not first or second person it will be in the third). Thus the eight syntactic properties become nine explanatory variables in the predictive model, as reported in Table 4.

Corpus	Predictor Coefficient
subject person: 1st	0.171
subject person: 2nd	1.280
plural subject	-0.300
pronoun subject	-0.470
zero subject	5.711
interrogative clause	2.139
past tense clause	-4.852
perfect aspect	-0.280
negated clause	-1.163
constant	-4.033

Table 4: Predictor coefficients for the presence of a zero auxiliary construction.

5.1 Model Evaluation

The logistic function is defined by:

$$f(Z) = \frac{1}{1 + e^{-z}} \quad (1)$$

The variable z is representative of the set of predictors and is defined by:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad (2)$$

where β_0, β_1, β_2 ... β_k are the regression coefficients of predictors x_1, x_2 .. x_k respectively. The predictors explored in this paper are encodings of the syntactic properties of the annotated sentences. The predictors and their encodings are indicated in Table 2.

The logistic function is constrained to values between 0 and 1 and represents the probability of membership of one of the two categories (zero auxiliary or auxiliary supplied). In our case an $f(z) > 0.5$ indicates that there is likely to be a zero auxiliary.

The logistic function defined by the coefficients in Table 4 is able to predict correct category membership for 96.9% of the sentences in the annotated corpus. All coefficients are highly significant to the logistic function ($p<0.001$) with the exception of perfect aspect and first person subject - which are both significant nevertheless ($p<0.05$).

For this model, positive coefficients indicate that the associated syntactic properties raise the probability of a zero auxiliary occurring. Large coefficients more strongly influence the probability of the zero auxiliary whereas near-zero coefficients have little influence. From the coefficients in Table 4 we see that the strongest predictor of a zero auxiliary is the occurrence of a zero subject (as in the utterance, *'leaving now.'*). An interrogative utterance is also a good candidate, as is the second person subject (e.g. *'you eating those olives?'*). However, a past tense utterance is an unlikely candidate for a zero auxiliary construction, as is a negated utterance.

6 Discussion — using the predictive model to aid parsing

As mentioned above, since parsers are trained on written data they can often display poor performance on text transcribed from the spoken domain. From the results of our corpus study we know that the zero auxiliary occurs in approximately 4.2% of progressive constructions in spoken language and we can extrapolate that it will occur in less than 1% (approximately 0.8%) of all progressive constructions in written language. A statistical parser trained on written language will therefore be prone to undergo parsing failure for every one in twenty-five progressive sentences. This is no insignificant problem, especially when it is remarked that the progressive is in high frequency usage (there are one thousand ING-forms featuring in progressive constructions for every one million words of sBNC) and that its use is known to be spreading (Leech et al, 2009).

Compounded with those parser breakdowns caused by other speech phenomena (for instance, repetition and elision), high numbers of parse failures on progressive constructions will render NLP accuracy on spoken language intolerable for any applications which rely on accurate parsing as a foundation. However, we have shown above that features of spoken language such as the zero auxiliary should not be thought of as errors or as unpredictable deviations from the written form, but rather can be considered to be consistent and predictable events. In this section we illustrate how our predictive model for zero auxiliary occurrence

$$(|relation|\ |head|\ |dependant|) \quad (3)$$
$$(|ncsubj|\ |play+ing:VVG|\ |you:PPY|) \quad (4)$$
$$(|obj|\ |play+ing:VVG|\ |what:DDQ|) \quad (5)$$
$$(|aux|\ |play+ing:VVG|\ |be+:VBR|) \quad (6)$$
$$(|arg|\ |play+ing:VVG|\ |you:PPY|) \quad (7)$$
$$(|relation|\ |verb:VVG|\ |dependant|) \quad (8)$$

Figure 1: Example grammatical relations from RASP.

may be integrated into a parser pipeline in order to aid the parsing of spoken language. In this way we build on the increasingly robust engineering of statistical NLP tools trained on written language by allowing them to adapt to the spoken domain on the basis of the linguistic study of speech phenomena.

In general the notion of 'parsing' an utterance involves a chain of several processes: utterance boundary detection, tokenization, part-of-speech tagging, and then parsing. We suggest that when it is known that the language to be parsed is from the spoken domain the pipeline of processes should be run in a SPEECH AWARE MODE. Extra functionality would be incorporated into each of the stages according to the findings of linguistic research into spoken language. In other work we have adapted the tokenization and tagging stages of the pipeline based on predictors that indicate when interjections (e.g. *'umm'*, *'err'* and *'ah'*) have been 'used' as punctuation or lexical items. We also incorporate intonation phrases as predictors for utterance boundary detection (Buttery and Caines: in preparation). Here, we augment the parsing stage of the pipeline by allowing an informed re-parse of utterances in which a parse failure is likely to have been caused by a zero auxiliary.

We present this section with reference to the specific mechanics and output formats of the RASP system but our algorithm is by no means parser specific and could be adapted for other parsers quite easily. Utterances parsed with RASP may be expressed as 'grammatical relations'. RASP's grammatical relations are theory-general, binary relations between lexical terms and are expressed in the form of head-dependancy relations as shown in (3), Figure 1.

Consider the utterance *'what are you playing?'*. When we parse this with RASP we get grammatical relations (4), (5) and (6) in Figure 1. The capital letter codes following the ':' symbols are part-of-speech tags (from the CLAWS-2 tagset (Garside, 1987)) which have been assigned to the lexical tokens by the tagger of the RASP system. Here *PPY* indicates the pronoun *'you'*; *VVG* indicates the ING-form of lexical verb; *VBR* indicates *'be'* in 3rd person present tense; and *DDQ* indicates a wh-determiner. The relation (4) tells us that *'you'* is the subject of *'playing'*; relation (5) tells us that *'what'* is taking the place of the object being played; and relation (6) tells us that there is an auxiliary relationship between *'are'* and *'playing'*. This is much as we would expect. However, if we try to parse *'what you playing?'* the parse fails. The single relation (4) is returned where ideally we would like both (4) and (5), as we did when the auxiliary was present.

For the utterance, *'you playing?'* RASP returns the under-specified grammatical relation (7) which is simply indicating that *'you'* is an argument of *'playing'* but not which type of argument (whether a subject, direct object, etc). Ideally we would like to retrieve at least (4) as we would have if we parsed the utterance *'are you playing?'*. For these examples, we shall consider the failure to identify the correct subject and object of the progressive verb to be a parsing failure.

We integrate the zero auxiliary predictive model with parsing technology to improve the parsing of zero auxiliaries in spoken language. Note that we use the RASP system but our algorithm is by no means parser specific. The only prerequisite is that the parser must be able to identify relations of some kind between the subject noun and ING-form (possibly via a parsing rule) and also be able extract values for the predictors (through either a rich tagset or from the identification of key speech tokens). The illustrative method we discuss here is integrated into the parsing pipeline in the event of a parse failure but there are several alternative methods that might also be considered.

For instance, by using the predictive model earlier in the parsing system pipeline a modified tagset could be used which updates the ING-form tag with a new tag to indicate that there is also a missing auxiliary. Another method might involve altering rule probabilities or adding extra parser rules so that parsing only has to occur once. Our other work in this area suggests that the final deci-

sion on where to add the spoken language modifications within the parsing pipeline will largely depend on the interaction of the phenomena in question with other speech phenomena.[6]

With the proviso that it is a preliminary integration of the predictive model into a parsing system, we propose the following algorithm for zero auxiliaries in spoken language. When 'speech aware mode' is activated, if we encounter a parse failure then we first check the part-of-speech tags of the utterance to ascertain if the sentence contains the ING-form requisite for a progressive construction:

- **IF no ING-form is found**: STOP. Our model predicts zero auxiliaries in progressive constructions—there is nothing more we can do with the input.
- **ELSE**: An ING-form is found. Extract all grammatical relations that were obtained by the parse which contained the ING-form in the head position (these would be grammatical relations that have the general format of (8) in Figure 1). We will refer to this set of grammatical relations as GRS.
  - **IF there is an auxiliary relation present in** GRS: STOP. If at least one of the extracted grammatical relations is an auxiliary relation, similar to (6) in Figure 1, an auxiliary is present—we do not have a zero auxiliary construction.[7]
  - **ELSE**: The utterance is a candidate for zero auxiliary.

Having determined a possible candidate for zero auxiliary we carry out the following steps:

1. Ascertain values for the zero auxiliary predictors (explained in more detail below).
2. Calculate the value of the logistic function $f(z)$ using the obtained predictor values with their coefficients (shown in Table 4).
3. If $f(z) > 0.5$, assume an auxiliary is missing.
4. Add the auxiliary to the sentence (choosing which auxiliary based on the predictor values—see below).
5. Re-parse the sentence.
6. Remove (or flag) the auxiliary grammatical relation from the newly obtained parser output.[8]

For step 1 above properties of the current utterance have to be obtained. The subject person, plural subject, zero subject and pronoun subject properties are ascertained by looking at the part-of-speech of the dependant noun/pronoun within any subject relations occurring in the set GRS (grammatical relations headed by the ING-form). Subject relations would look similar to (4) in Figure 1. If there is no subject grammatical relation, any underspecified 'arg' relation (such as (7) in Figure 1) are considered. If neither of these relations are present in GRS then a zero subject is inferred. The person and plurality of the subject noun is encoded within its CLAWS2 part-of-speech tag. For instance, a *PPHS1* tag, which is used to indicate *'him'* or *'her'* would tell us we have a third person, singular pronoun.[9]

The other properties are all ascertained by the presence or absence of a token within the utterance: interrogative property is inferred when the utterance ends with a question mark; the negation property when either *'not'* or *'n't'* (which are tagged *XX*) is present; the perfect is inferred from the presence of the word *'been'*; and past tense is ascertained from a set of temporal marker lexical items (e.g. *'yesterday*, *'before'*). Once extracted the properties are encoded as shown in Table 4 for use as the predictor values in the logistic function.

In order to select the correct auxiliary and location for insertion in step 4 the utterance values are consulted. For instance, an interrogative utterance in the present tense, not in perfect aspect, with a second person singular subject will require insertion of the auxiliary *'are'* after the subject. A zero subject zero auxiliary, on the other hand, requires restoration of both subject and auxiliary. Where a question mark indicates it has been used in an interrogative clause the subject is assumed to be sec-

[6] Although, another major consideration is the overall computational efficiency of the parsing system.

[7] This step is actually subtly more complicated—auxiliary relations involving *'been'* are allowed to be present in GRS (this allows us to capture zero auxiliaries in the perfect such as *'been coming here long?'*) but if there is any other auxiliary relation present in GRS then we STOP here.

[8] We also remove (or flag) the subject relation in cases where a subject also had to be added in step 4. This would occur when the original utterance exhibited a zero subject.

[9] All common nouns are assumed to be 3rd person and all instances of *'you'* were considered to be singular (as was the case during corpus annotation).

ond person - as is the case in most questions - and so the auxiliary-subject combination *'are you'* is restored before the ING-form. Without a question mark, the clause is assumed to be declarative and so the first person singular subject-auxiliary combination *'I am'* is restored before the ING-form [10].

We withheld 10% of the zero auxiliary corpus for test purposes. The integration of the predictive model into the parser allowed us to successfully parse 31.4% of previously unparsable zero-auxilaries. On cleaned spoken transcripts (i.e. with speech phenomena other than the zero auxiliary, such as repetitions, removed) this algorithm allows us to retrieve the correct subject-object relations for an extra 1238 utterances within our annotated corpus (which again accounts for approximately one third of the previously unparsable zero-auxilaries). This is a significant step forward for any applications building on top of a parsing infrastructure.

7 Conclusion

We have shown how awareness of a specific linguistic phenomenon enables improvements in NLP techniques. The zero auxiliary is mainly a feature of spoken language and so is not on the whole handled successfully by existing parsers, trained as they are on written data. As a solution, rather than proposing the construction of new models specifically designed for spoken language, thereby doing away with all previous work on NLP tools and starting again from scratch, we demonstrated how new training data from a spoken source could be applied to an existing parser - RASP. We designed a predictive model of zero auxiliary occurrence based on logistic regression with nine syntactic variables. The data came from an annotated corpus of 93,253 progressive constructions which showed zero auxiliary frequency to be 4.2%. Without this new predictive information in the parser, the status quo would continue whereby one in twenty-five progressive constructions would continue to be mis-parsed. We found that instead the noise was regular and could be modelled, and we illustrated how this specific linguistic data could be integrated into existing NLP technology. This is a case study of one specific linguistic phenomenon. Our belief is that other such spoken language phenomena can be modelled in the same way, given an appropriate corpus resource, accurate annotation and implementation into a parser.

By running in a 'speech aware mode' which supplements existing parsing architecture we benefit from the training that has already been undertaken on a large scale based on written data and complement it with specialized and predictable linguistic properties of speech. Ideally, we would like to train an entire parsing system on spoken language but until spoken corpora become more readily available this is not a practical option: the resulting parser would suffer greatly from data sparsity issues. Frustratingly, there is a circular problem in generating corpora of an appropriate size for training since until highly accurate models for spoken language are built we can not expect speech-to-text systems to provide highly accurate transcripts. But to build these highly accurate models of spoken language in the first place a large amount of data is required. Augmenting the existing statistical NLP tools trained on written language with specialized linguistic knowledge from the spoken domain is a pragmatic short-term fix for this problem.

We should note that tailoring parsers to deal with spoken language is by no means unheard of: the RASP system itself, for example (which parses using a probabilistic context-free grammar), already has several rules in its grammar which are more appropriate for parsing spoken language. However, use of these rules can contribute to much over-generation and complexity in the parse forest (the parser internal structure which holds all the possible parses for an utterance). In consequence, the specialized rules have to be expertly selected or deselected when configuring the parser. This work - and our research program as a whole - would instead allow parser configuration decisions and algorithmic adaptions to be made non-expertly and on-the-fly when running in 'speech aware mode'. All rule activations and algorithm adaptions would be made based on predictions constructed from expert linguistic analysis of the spoken domain.

Acknowledgements

This work was supported by the AHRC. We thank three anonymous reviewers for their comments, Andrew Rice and Barbara Jones.

[10] A sample of one hundred zero subject declarative zero auxiliaries indicates that the first person singular is the appropriate subject type to restore on 60% of occasions.

References

Gisle Andersen. 1995. Omission of the primary verbs BE and HAVE in London teenage speech - a sociolinguistic study. Hovedfag thesis, University of Bergen, Norway.

Michael Barlow and Suzanne Kemmer. 2000. *Usage-based models of language*. Chicago: CSLI.

Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.). 2003. *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Ted Briscoe, John Carroll and Rebecca Watson. 2007. The second release of the RASP system. *Proceedings of the COLING/ACL on Interactive presentation sessions*, July 17-18, 2006, Sydney, Australia.

The British National Corpus, version 3. 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Roger Brown. 1973. *A First Language: the early stages*. London: George Allen and Unwin.

Paula Buttery and Andrew Caines. In preparation. *An Empirical Approach to First Language Acquisition*. Cambridge: Cambridge University Press.

Joan Bybee and Paul Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Andrew Caines. 2010. *You talking to me?* Zero auxiliary constructions in British English. Ph.D thesis, University of Cambridge.

Frantisek Cermák. 2009. Spoken corpora design. Their constitutive parameters. *International Journal of Corpus Linguistics* 14: 113-123.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Roger Garside. 1987. The CLAWS Word-tagging System. In: Roger Garside, Geoffrey Leech and Geoffrey Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Michael A. K. Halliday. 1994. Spoken and Written Modes of Meaning. In: David Graddol and Oliver Boyd-Barrett (eds.), *Media Texts: Authors and Readers*. Clevedon: Multilingual Matters.

Michael Haugh. 2009. Designing a Multimodal Spoken Component of the Australian National Corpus. In: Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters (eds.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*. Somerville, MA: Cascadilla Proceedings Project.

William Labov. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45: 715-762.

Geoffrey Leech, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English: a grammatical study*. Cambridge: Cambridge University Press.

Elena Lieven, Heike Behrens, Jennifer Speares and Michael Tomasello. 2003. Early syntactic creativity: a usage-based approach. *Journal of Child Language* 30: 333-370.

Luigi Rizzi. 1993/1994. Some notes on linguistic theory and language development: The case of root infinitives. *Language Acquisition* 3: 371-393.

Geoffrey Sampson. 2001. *Empirical Linguistics*. London: Continuum.

Anna Theakston, Elena Lieven, Julian Pine and Caroline Rowland. 2005. The acquisition of auxiliary syntax: BE and HAVE. *Cognitive Linguistics* 16: 247-277.

Harry Tily, Susanne Gahl, Inbal Arnon, Neal Snider, Anubha Kothari and Joan Bresnan. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1: 147-165.

Kenneth Wexler. 1994. Optional Infinitives, head movement and the economy of derivations. In: David Lightfoot and Norbert Hornstein (eds.), *Verb Movement*. Cambridge: Cambridge University Press.

Stephen Wilson. 2003. Lexically specific constructions in the acquisition of inflection in English. *Journal of Child Language* 30: 75-115.